# Homework 1

CSE 541: Interactive Learning
Instructor: Kevin Jamieson
Due 11:59 PM on April 15, 2024

**Probability**

Concentration inequalities are at the heart of most arguments in statistical learning theory and bandits. Refer to [1] for more details.

1.1 (Markov's Inequality) Let $X$ be a positive random variable. Prove that $\mathbb{P}(X > \lambda) \leq \frac{\mathbb{E}[X]}{\lambda}$.

1.2 (Jensen's Inequalty) Let $X$ be a random vector in $\mathbb{R}^d$ and let $\phi : \mathbb{R}^d \to \mathbb{R}$ be convex. Then $\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)]$. Show this inequality for the special case when $X$ has discrete support. That is, for $p_i \geq 0$ and $\sum_{i=1}^n p_i = 1$, and $(x_1, \ldots, x_n) \in \mathbb{R}^n$ show that $\phi(\sum_{i=1}^n p_i x_i) \leq \sum_{i=1}^n p_i \phi(x_i)$.

1.3 (Sub-additivity of sub-Gaussian) For $i = 1, \ldots, n$ assume $X_i$ is an independent random variable with $\mathbb{E}[\exp(\lambda(X_i - \mathbb{E}[X_i])] \leq \exp(\lambda^2 \sigma_i^2/2)$. If $Z = \sum_{i=1}^n X_i$ find $a \in \mathbb{R}$ and $b \geq 0$ such that $\mathbb{E}[\exp(\lambda(Z-a))] \leq \exp(\lambda^2 b/2)$.

1.4 (Maximal inequality) For $i = 1, \ldots, n$ let each $X_i$ be an independent, random variable that satisfies $\mathbb{E}[\exp(\lambda X_i)] \leq \exp(\sigma_i^2 \lambda^2/2)$ for all $\lambda > 0$. Show that $\mathbb{E}[\max_{i=1,\ldots,n} X_i] \leq \sqrt{8 \max_{i=1,\ldots,n} \sigma_i^2 \log(n)}$. Hint[1]. If $\sigma_1 \gg \sigma_2 = \cdots = \sigma_n$ how would you expect $\mathbb{E}[\max_{i=1,\ldots,n} X_i]$ to behave (intuitive justification is enough)?

**The Upper Confidence Bound Algorithm.**

Consider the following algorithm for the multi-armed bandit problem.

**Algorithm 1**: UCB

**Input:** Time horizon $T$, 1-subGaussian arm distributions $P_1, \cdots, P_n$ with unknown means $\mu_1, \cdots, \mu_n$ such that $\mathbb{E}_{X \sim P_i}[X] = \mu_i$

**Initialize:** Let $T_i(t)$ denote the number of times arm $i$ has been pulled up to (inclusive) time $t$ and let $T_i = T_i(T)$. Pull each arm once.

**for:** $t = n+1, \cdots, T$

Pull arm $I_t = \arg\max_{i=1,\cdots,n} \widehat{\mu}_{i,T_{i(t-1)}} + \sqrt{\frac{2\log(2nT^2)}{T_i(t-1)}}$ and observe draw from $P_i$

Let $\widehat{\mu}_{i,T_i(t)}$ be the empirical mean of the first $T_i(t)$ pulls.

In the following exercises, we will compute the regret of the UCB algorithm and show it matches the regret bound from lecture. Without loss of generality, assume that the best arm is $\mu_1$. For any $i \in [n]$, define the *sub-optimality gap* $\Delta_i = \mu_1 - \mu_i$. Define the regret at time $T$ as $R_T = \mathbb{E}[\sum_{t=1}^T \mu^* - \mu_{I_t}] = \sum_{i=1}^n \Delta_i \mathbb{E}[T_i]$.

2.1 Consider the event

$$\mathcal{E} = \bigcap_{i \in [n]} \bigcap_{s \leq T} \left\{ |\widehat{\mu}_{i,s} - \mu_i| \leq \sqrt{\frac{2\log(2nT^2)}{s}} \right\}.$$

Show that $\mathbb{P}(\mathcal{E}) \geq 1 - \frac{1}{T}$.

2.2 On event $\mathcal{E}$, show that $T_i \leq 1 + \frac{8\log(2nT^2)}{\Delta_i^2}$ for $i \neq 1$.

2.3 Show that $\mathbb{E}[T_i] \leq \frac{8\log(2nT^2)}{\Delta_i^2} + 2$. When $n \leq T$, conclude by showing that $R_T \leq \sum_{i=2}^n \left( \frac{24\log(2T)}{\Delta_i} + 2\Delta_i \right)$.

[1] Apply Jensen's inequality to the identity $\mathbb{E}[\max_i X_i] = \frac{1}{\lambda}\log(\exp(\lambda \mathbb{E}[\max_i X_i]))$ for any $\lambda > 0$

**Thompson Sampling**.
We consider the following Bayesian setting. Consider $n$ arms and let $p_0$ be an $n$-dimensional prior distribution over $[-1,1]^n$ such that $\theta^* \sim p_0$ is drawn before the start of the game (e.g., $p_0$ is uniform over $[-1,1]^n$). At any time $t$, when we pull arm $i \in [n]$ we observe a random variable $X_{i,t} \in [-1,1]$ where $\mathbb{E}[X_{i,t}] = \theta_i^*$.

**Algorithm 1**: Thompson Sampling

**Input:** Time horizon $T$
Assume the prior distribution $p_0$ over $\mathbb{R}^n$ is known and that $\theta^* \sim p_0$ (so that $\theta^* \in \mathbb{R}^n$). Assume each arm shares the same conditional likelihood function such that an observation $X$ from arm $i$ follows $X \sim f(\cdot|\theta_i^*)$ (e.g., $X \sim \mathcal{N}(\theta_i^*, 1)$). Let $p_t(\theta|I_1, X_{I_1,1}, \cdots, I_t, X_{I_t,t}) \propto \prod_{s=1}^t f(X_{I_s,s}|\theta_{I_s})p_0(\theta)$ be the posterior distribution on $\theta^*$ at time $t$.
**for:** $t = 1, \cdots, T$
Sample $\theta^{(t)} \sim p_{t-1}$ (Note: $\theta^{(t)} \in \mathbb{R}^n$)
Pull arm $I_t = \arg\max_{i \le n} \theta_i^{(t)}$ to observe $X_{I_t,t}$
Compute exact posterior update $p_t$

Denote the $\sigma$-algebra generated by the observations at time $t$ by $\mathcal{F}_t = \sigma(I_1, X_{I_1,1}, \cdots, I_t, X_{I_t,t})$ (if you are unfamiliar with $\sigma$-algebras, don't worry too much - conditioning on the $\sigma$-algebra just means conditioning on the choices of arms and the rewards observed). For any event $A \in \mathcal{F}_t$ let $\mathbb{P}_t(A)$ denote the probability under $\mathcal{F}_t$. The *Bayesian Regret* of an algorithm is

$$\begin{aligned} BR_T &= \mathbb{E}_{\theta^* \sim p_0}\left[\sum_{t=1}^T \max_{i=1,\ldots,n} \theta_i^* - \theta_{I_t}^*\right] \\ &= \mathbb{E}_{\theta^* \sim p_0}\left[\mathbb{E}_{I_t}\left[\sum_{t=1}^T \max_{i=1,\ldots,n} \theta_i - \theta_{I_t}\Big|\theta^* = \theta\right]\right] \\ &= \mathbb{E}_{\theta^* \sim p_0}\left[\mathbb{E}_{I_t}\left[\sum_{t=1}^T \mathbb{E}_{I_t}[\max_{i=1,\ldots,n} \theta_i - \theta_{I_t}|\mathcal{F}_t, \theta]\Big|\theta\right]\right] \end{aligned}$$

Assume that expectations, if not explicitly specified, are with respect to all randomness including $\theta^* \sim p_0$, $I_1, \ldots, I_T$, and observations that contribute to $\mathcal{F}_t$.

3.1 On a given run of the algorithm, let $\widehat{\theta}_{i,s}$ denote the empirical mean of the first $s$ pulls from arm $i$, note that $\mathbb{E}[\widehat{\theta}_{i,s}] = \theta_i^*$. Let the good event be

$$\mathcal{E} = \bigcap_{i \in [n]} \bigcap_{t \le T} \left\{ |\widehat{\theta}_{i,t} - \theta_i^*| \le \sqrt{\frac{2\log(2/\delta)}{t}} \right\}.$$

Show that $\mathbb{P}(\mathcal{E}^c) \le nT\delta$.

3.2 (Key idea.) Argue that for all $i \in [n]$ that $\mathbb{P}(\arg\max_{j=1,\ldots,n} \theta_j^* = i|\mathcal{F}_{t-1}) = \mathbb{P}(I_t = i|\mathcal{F}_{t-1})$. Note that the probability on the right hand side is over the posterior distribution over $\theta_*$ only, whereas the left-hand-side is also over the randomness of $I_t$.

3.3 Define $U_t(i) = \min\{1, \widehat{\theta}_{i,T_i(t)} + \sqrt{\frac{2\log(2/\delta)}{T_i(t)}}\}$. For any $\theta \in \mathbb{R}^d$ define $i_*(\theta) = \arg\max_i \theta_i$. Show that $\mathbb{E}_{\theta^* \sim p_0}[\mathbb{E}_{I_t}[\theta^*_{i^*(\theta^*)} - \theta^*_{I_t}|\mathcal{F}_{t-1}]] = \mathbb{E}_{\theta^* \sim p_0}[\theta^*_{i^*(\theta^*)} - U_t(i^*(\theta^*))] + \mathbb{E}_{\theta^* \sim p_0}[\mathbb{E}_{I_t}[U_t(I_t) - \theta^*_{I_t}|\mathcal{F}_{t-1}]]$. Conclude that $BR_T = \mathbb{E}_{\theta^* \sim p_0}[\sum_{t=1}^T \theta^*_{i^*(\theta^*)} - U_t(i^*(\theta^*)) + \sum_{t=1}^T \mathbb{E}_{I_t}[U_t(I_t) - \theta^*_{I_t}|\mathcal{F}_{t-1}]]$. Hint[2].

[2]Tower rule of expectation.

3.4 Show that $BR_T \leq 4\delta nT^2 + \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\{\mathcal{E}\}\left(\sum_{t=1}^{T} U_t(I_t) - \theta^*_{I_t}\right)\middle|\theta^*\right]\right] \leq O(\delta nT^2 + \sqrt{Tn\log(1/\delta)})$. Hint[3]

3.5 Make an appropriate choice of $\delta$ and state a final regret bound.

In general, giving frequentist bounds on the regret is significantly more difficult. We refer the interested reader to [2] and the tutorial [3] for more details. This exercise is motivated by [4]

**Algorithm 1**: Explore-then-Commit

**Input:** Time horizon $T$, $m \in \mathbb{N}$, 1-sub-Gaussian arm distributions $P_1, \cdots, P_n$ with unknown means $\mu_1, \cdots, \mu_n$
**for:** $t = 1, \cdots, T$
If $t \leq mn$, choose $I_t = (t \mod n) + 1$
Else, $I_t = \arg\max_i \widehat{\mu}_{i,m}$

**Empirical Experiments**
Implement UCB, Thompson Sampling (TS), and Explore-then-Commit (ETC). Let $P_i = \mathcal{N}(\mu_i, 1)$ for $i = 1, \ldots, n$. For Thompson sampling, define the prior for the $i$th arm as $\mathcal{N}(0, 1)$ and the likelihood function as $f(\cdot|\mu_i) = P_i$.

4.1 Let $n = 10$ and $\mu_1 = 0.1$ and $\mu_i = 0$ for $i > 1$. On a single plot, for an appropriately large $T$ to see expected effects, plot the regret for the UCB, TS, and ETC for several values of $m$.

4.2 Let $n = 40$ and $\mu_1 = 1$ and $\mu_i = 1 - 1/\sqrt{i-1}$ for $i > 1$. On a single plot, for an appropriately large $T$ to see expected effects, plot the regret for the UCB, TS, and ETC for several values of $m$.

# References


[1] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[2] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1, 2012.

[3] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

[4] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.


[3] Apply Jensen's to $\sum_{i=1}^{n} \sqrt{T_i}$.